# Episogram: Visual Summarization of Egocentric Social Interactions

Nan Cao, Yu-Ru Lin, Fan Du, Dashun Wang

**Abstract**—Visualizing social interaction data has been of booming interest as the recent availability of social traces, ranging from the conversations left in social media to groups' collaborations archived in publications. The key challenges of visualizing social interaction data including the difficulties of (1) understanding the general structure of social interactions and (2) representing the data in the context of different user activities for revealing different behavior patterns. In this paper, we present, Episogram, for visualizing social interaction data. Our design is based on an anatomy of social interaction process in which the actors and objects involved can be formally represented as a time-varying tripartite network. In Episogram, we display and aggregate such tripartite networks along multiple temporal dimensions, from different actors' egocentric perspectives. We show the effectiveness of the proposed technique via case studies and user studies. The results indicate that our design provides non-trivial insights from social interaction data.

**Index Terms**—Social Interactions, Social Media Visualization, Visual Summarization, Information Visualization.

## 1 Introduction

Social interaction refers to a "dynamic, changing sequence of social actions between individuals (or groups) who modify their actions and reactions due to the actions by their interaction partner(s)"[1]. Nowadays, datasets that archive the social interactions of individuals have become increasingly available. Examples include the content generated by hundreds of millions of users on social media such as Twitter, the communications and transactions recorded in emails and instant messages, and publications that documented the collaboration among authors. These social traces provide a proliferation of opportunities for understanding social interactions, which are considered to be important. For example, understanding the common features of users' communication activities helps analysts identify their common behaviors, thus facilitating the detection of anomaly users, which is a serious need in social security. However, understanding these data is not an easy task given the complexity of the datasets (unstructured, dynamic, and heterogeneous) and the variations of different types of social interactions in various application domains.

Data visualization enables understanding complex data through intuitive representations, facilitating data interpretation and summarization. However, several challenges exist in visualizing the social interaction data. First, the activities occurred during the social interactions (e.g., posting or retweeting tweets) provide necessary contexts for understanding the meaning of the interactions [1]. Therefore, an efficient visualization should be able to display and capture such rich-context social interactions with a simple and an integrative visual design. Designing such a visualization is non-trivial. Second, designing a visualization for capturing the temporal patterns (e.g., frequency and duration of the social interaction process), content patterns (e.g., the topics around which the interaction occurred), and behavior patterns (e.g., how a user post or retweet in Twitter) is important for revealing the insight of the social interaction data, but it is a hard to achieve. Furthermore, there is lack of understanding of the common structure in social interaction processes, which is the key for overcome the above challenges.

In this paper, we introduce a novel visualization design, "Episogram", for visualizing social interaction data (Fig. 1). The key contributions in this work include: (1) We provide an in-depth analysis of the key elements and structure of the social interaction process. Followed by this analysis, we introduce a directed tripartite network data model that can capture essential social interaction information in generalized social contexts. (2) We extend the Andrienko task model [2] to characterize different levels of user tasks in seeking information in social interaction data. Based on this task requirement, we propose a novel egocentric representation for visualizing individuals' interaction histories. The egocentric representation conveys two types of roles an individual may play during an interaction process, as an *initiator* or as a *responder*, with two types of layouts for effectively identifying and comparing interaction patterns.

## 2 Related Work

Episogram extends prior work in visualization of time-oriented data. A summarization of the techniques in this area can be found in [2]. We compare our work with the most related existing designs via a controlled user study as described in section 6. Here, we focus on comparing our work with those visualizations designed for summarizing social activities in order to understand the design limitations in existing work.

*Nan Cao is with IBM T. J. Watson Research Center. E-mail: nan.cao@gmail.com.*
*Yu-Ru Lin is with University of Pittsburgh. E-mail: yurulin@pitt.edu.*
*Fan Du is with University of Maryland. E-mail: fan@cs.umd.edu.*
*Dashun Wang is with Pennsylvania State University. E-mail: dwang@ist.psu.edu.*

1. Interaction. http://en.wikipedia.org/wiki/Interaction

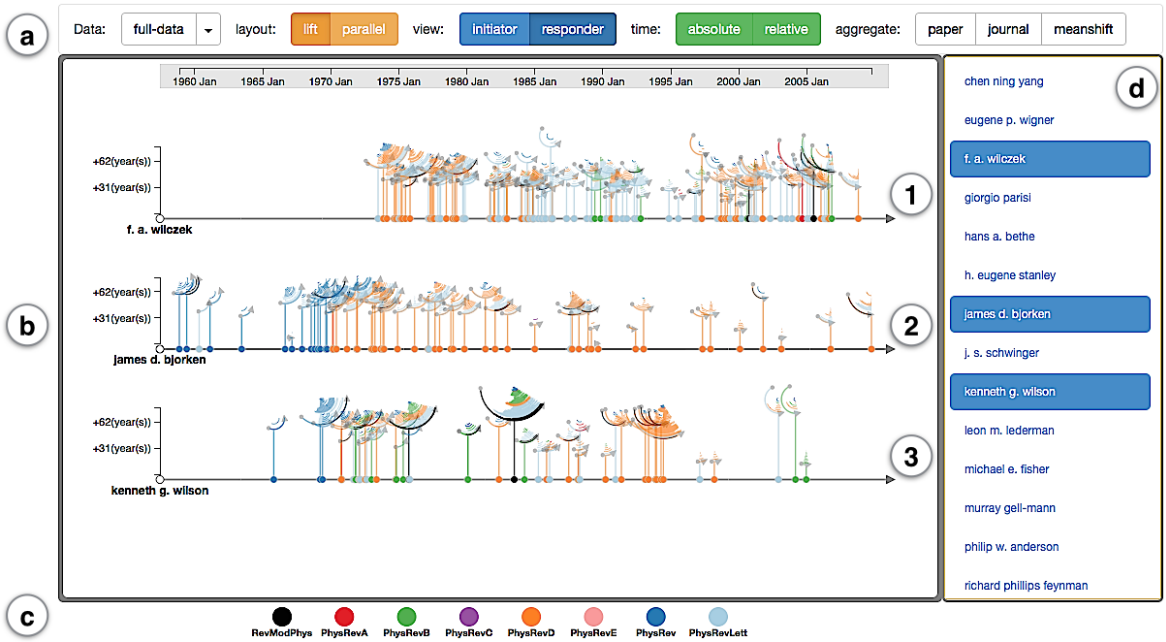


Fig. 1. Episogram enable users to explore and compare social interaction data from egocentric views. In this case, the social interactions of the social actors (the three scholars in Physics) are visualized along the timelines. The social interaction events, including publishing papers (represented as vertical bars) and receiving citations (represented as crescent shapes on top of the vertical bars) were scattered according to when the events occurred. The scholar (1), Dr. Wilczek, has constantly published since early '70, and most of his papers were published and got cited in journals Physical Review D and Physical Review Letters (differentiated in colors). The scholar (2), Dr. Bjorken, was very productive between 1965 and 1990. His renowned work began with publications in Physical Review and later he published more on Physical Review D. The scholar (3), Dr. Wilson, has an interesting trajectory. He had two productive periods, between 1970 and 1975, and between 1990 and 1995, respectively. Also interestingly, his most cited work was published in 1983, in Reviews of Modern Physics. The visualization is generated based on a complete collection of papers published by Physical Review, as well as citations among them. By visualizing the history of scholars' social interactions with the event contexts (e.g., journals), this figure allows comparing the productivity and impact of the three scholars in the fields. The label annotations from (a) to (d) correspond to major UI components which will be described in detail in the paper.

There has been work aiming at providing visual summarization of wide-ranging activities. Ogawa *et al.* [3] represented the transition of email exchange in open source software projects through Sankey diagrams [4]. Some work employed glyph-based designs. The goal is to identifiable summarization of different activities. For example Erbacher *et al.* [5] introduced a radial glyph that summarizes a web server's activity of connecting to other severs over time. Anemone [6] introduced a glyph showing the statistical information of users' visiting a web page. These designs summarized the activities at a given time point as a glyph and the changes of activities were displayed via animation. PeopleGarden [7] introduced a flower shaped glyph for summarizing a user's aggregated interaction histories in a discussion group. The flower glyphs of different users are randomly placed in a display area called "garden". Although it summarized users' interactions, all the details such as "when did who involved in an interaction" are unavailable from such visualization. These designs may be useful in providing a snapshot view or an aggregated view of interaction history, but they are not effective for identifying or comparing temporal patterns from the data. HistoryFlow [8] introduced a stacked flow visualization that displayed collaborations of the users who edited on the same Wikipedia page. This visualization allowed users to compare interaction (i.e. co-editing a page) patterns within a limited interaction context (a single Wiki page). It is thus difficult to extend the design to a more general setting or compare the change of interaction context over time.

## 3 Data Model and Terminology

In this section, we identify the key elements and structure of the social interaction process, which provides a basis for the terminology and data model that will be used in our visualization design.

In our day-to-day social experience, social interactions form the basis of social relations. A social interaction can be any relationship between two or more individuals that consists of a sequence of *interaction events*. It is an essential component that drives various communication technologies – in social media like Twitter, interactions are manifested through "tweeting" (a user posts a tweet) and "replying" or "retweeting" (users rebroadcast a tweet posted by others). Social interactions commonly involve *social objects*, i.e., the content around which conversation happens [9]. Examples of social objects include emails (in email exchanges), tweets (in Twitter communications), papers (in co-authorship), and various types of artifacts. A social object connects people with shared interests in a social interaction. There are two types of roles an individual may play during an interaction process: an *initiator* who initiates the interaction by creating an social object, and a *responder* who responds by acting on the social object created by the initiator. For example, suppose Alice and Bob are two users interacting with each other on Twitter, and suppose Alice posts a tweet on which

Bob retweets. In this case, Alice is an initiator, Bob is a responder, and the tweet is a social object.

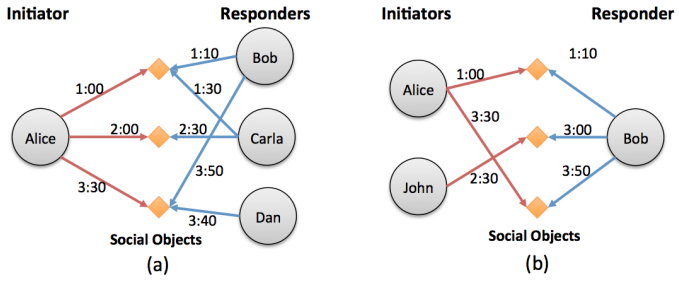


Fig. 2. Data model for social interactions. (a) The initiator-centric model. (b) The responder-centric model.

We introduce a *directed tripartite network* model to represent the key elements and structure of an interaction process. As shown in Fig. 2, initiators and responders are denoted as two types of nodes on either left or right side, with social objects as the third type of nodes connecting both initiators and responders. Actions, including initiating and responding with respect to a social object, are denoted as directed edges pointing to the social objects, with timestamp indicating the time when the action occurred. For example, in Fig. 2(a), Alice posted three tweets (social objects) at 1:00, 2:00, and 3:30. The first tweet was retweeted by Bob and Carla, the second tweet was retweeted by Carla, and the third tweet was retweeted by Bob and Dan. In this network, Alice is an initiator with actions (posting tweets) represented as red edges, the tweets (orange diamond nodes) are social objects, and Bob, Carla and Dan are three responders whose actions (retweeting) are represented as blue edges. We call this an initiator-centric model since Alice (the initiator) is of the central interest of all actions shown in this network. In contrast, Fig. 2(b) shows a responder-centric model where the responder is of central interest. In this network, Bob (responder) retweeted three tweets posted by Alice and John (initiators). Note that an individual can be both an initiator and a responder at the same time, but in an initiator-centric (responder-centric) model, his/her responding (initiating) actions are omitted.

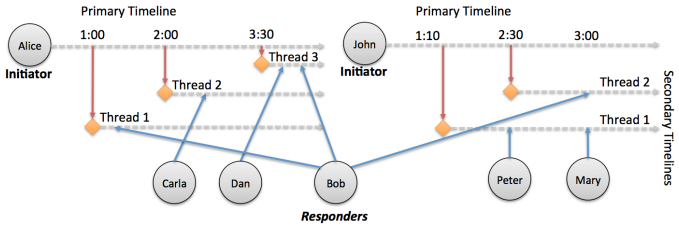


Fig. 3. The timeline representation of the model for interactions.

Putting together, social interactions involving a set of initiators and responders can be combined to emphasize the temporal relationship of the interaction events. As shown in Fig. 3, interaction events initiated by the initiators, Alice and John, are carried on the *primary timelines*. Each of the social objects created through these initiating actions can be acted upon by different responders. We call the initiating event and the subsequent responding events associated with the same object an activity *thread*. The subsequent responding events on an activity thread are carried on a *secondary timeline* (as opposed to a primary timeline due to its dependency on the creation of thread). For example, Alice posted three tweets which are starting points of three activity threads. The retweeting events (e.g., Dan and Bobs retweeting on the third tweet) are carried on the secondary timelines associated with each of the threads.

An initiator (e.g., Alice in Fig. 3) can generate multiple activity threads through creating different social objects, and a responder (e.g., Bob in Fig. 3) can connect to multiple threads through responding to different social objects created by the same or different initiators. Based on the above anatomy, the problem of visualizing social interaction history can be thus approached by creating a tool for exploring the various kinds of temporal relationships contained in the connected tripartite networks as shown in Fig. 3.

## 4 Visualizing the Social Interaction Data

In this section, we present our visualization design as an approach for visualizing social interaction data.

### 4.1 Design Goals and Tasks

The overall goal of our visualization design is to help users to gain insights from the social interaction data via data exploration. We decompose this goal into a set of tasks that users might seek to answer. We extend the Andrienko task model [2] by characterizing three levels of user tasks in seeking information in social interaction data.

**Elementary tasks.** Elementary tasks address individual data elements. In the context of visualizing interaction history, the user tasks include:

- **T1** (*look up*): How (through what social object) did actor $A$ interact with actor $B$ at certain time $T$? (*direct lookup*) When did actor $A$ interact with actor $B$? (*inverse lookup*)
- **T2** (*comparison and relation seeking*): Compare how actor $A$ interacted with actor $B$ differently with actor $C$. (*direct comparison*) When actor $A$ initiated an interaction by creating a social object, did actor $B$ respond before or after others? (*inverse comparison*) When did actor $B$ respond to actor $A$ quicker than others? (*relation seeking*)

**Synoptic tasks.** Synoptic tasks involve a general view of data. Here, the user tasks include:

- **T3** (*pattern identification and search*). What was the frequency of interaction between actor $A$ and others during certain time $T$? (*pattern identification*) When did actor $A$ interact with others frequently? (*pattern search*)
- **T4** (*pattern comparison*). Compare the interaction frequency between actor $A$ and others during time $T_1$ and time $T_2$. How does others respond to actor $A$ during $T_1$ and $T_2$?

**Higher-level synoptic tasks.** One of the key motivations for visualizing social interaction data is to characterize individuals' social behavior and further gain insights from

comparing how people interact with others might affect their life outcomes (e.g., work productivity or career path). Hence, we identify higher-level (i.e., more abstract) synoptic tasks based on the identification, search and comparison of patterns about individual social actors.

**T5** (*actor pattern identification and search*). Did actor $A$'s interaction with others persist over a long period of time or concentrate during a certain time? When did actor $A$'s interaction with others suddenly increase?

**T6** (*actor pattern comparison*). How did actor $A$'s interaction with others different from actor $B$? Was actor $A$ more active (in terms of initiating an interaction) than actor $B$? Was actor $A$ more responsive (in terms of responding others' interaction) than actor $B$?

We designed Episogram iteratively based on the above tasks by closely working with an expert with background in computational social science. A weekly discussion lasted for about 1.5 months was hold for us to develop an effective visual design. In each design iteration, several design choices were proposed and drawn manually based on a small set of toy data for illustrating the concept. The expert evaluated their effectiveness, identified their limitations, and provided design suggestions for improvement by applying them to solve the aforementioned tasks. Finally, two designs, Gantt Chart and the one proposed in the this paper (i.e., Episogram), were considered to be the most effective among all other design choices. We conducted a formal controlled user study (section 6) to compare these two designs. The results illustrated several significant benefits of the second design which will be introduced in the next section.

### 4.2 Visualization Design

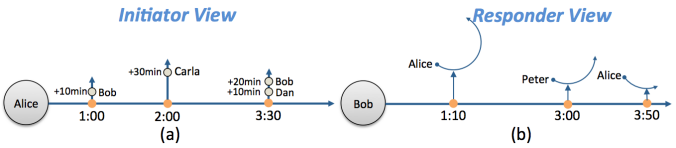


Fig. 4. Design overview of Episogram based on the combination of two different views (a) initiator view and (b) responder view.

Our design seeks to help users find answers for the above-mentioned tasks from the temporal social interaction data as illustrated in Fig. 3. We propose an *egocenric* representation – to focus on each individual's interaction at a time, based on the role he or she plays in the social interactions. In particular, this egocentric data can be shown in (1) *initiator view*: when and how the individual initiated interaction events by creating social objects, and (2) *responder view*: when and how the individual participated in interactions through responding with respect to social objects created by others. Fig. 4(a) and (b) show the two views extracted from the networks in Fig. 3.

In Fig. 4(a), the primary timeline carries time points when Alice posted tweets. Each activity thread (Fig. 5(a)), represented as the vertical line, interacts with the primary timeline at the time point $t$ – the time when the corresponding social object (shown as a circle at the intersection) is created. All subsequent responding events with respect

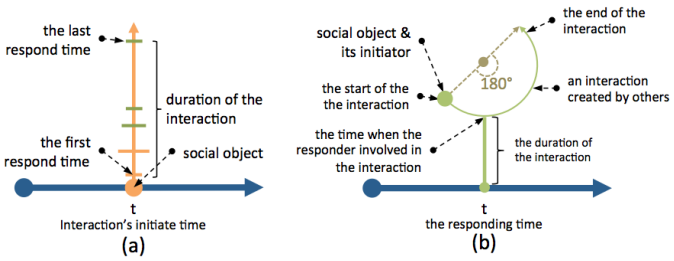


Fig. 5. Activity thread in (a) initiator view and (b) responder view.

to the social object are marked on the vertical line with intersections indicating when the responding events occur. The length of the vertical line depends on the lag of the last responding events on the thread.

In the responder view shown in Fig. 4(b), the primary timeline carries time points when Bob retweeted others' tweets. Each interaction thread (Fig. 5(b)) is represented as a crescent-shape curved arrow lifted by a vertical line. The crescent shape begins with a circle (representing the corresponding social object) indicating the creation of the thread – the time when the social object is created. The crescent shape ranges from 0 to 180 degrees indicating the relative duration of the corresponding social interaction thread. A 180-degree crescent shape represents the longest duration of the activity thread in the dataset. The length of the vertical line double encodes the duration of the responded thread. The intersection between the crescent shape and the vertical line shows when the responder participates in the activity thread, e.g., the responder's retweeting time of a tweet. Hence, the orientation of the crescent shape reflects how early or late a responder participates in the activity thread.

In both views, color and size can be used in thread to represent additional data attributes such as the sentiment and the number of retweets of a tweet. In addition, arranging the vertical thread lines parallelly together with their start points connecting to the primary timeline facilitates a fast comparison of durations of different threads, thus enabling an easy detection of influential threads.

### 4.3 Threads Aggregation

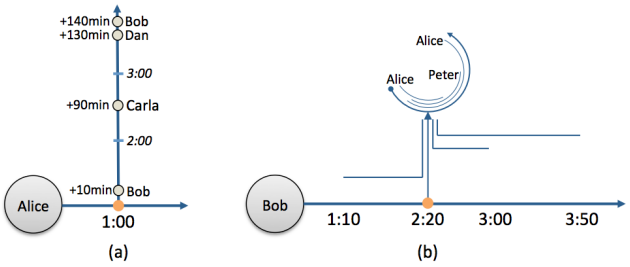


Fig. 6. Visualizing a thread cluster in (a) initiator view and (b) responder. These two examples show the aggregation of the activity threads in Fig. 4 (c) and (d) respectively.

We introduce a *thread aggregation* design to reduce the visual clutter caused by dense social interaction events and help detect potential events in social interactions.

In the initiator view, a cluster of threads can be visualized by directly merging multiple threads into the same

vertical line. This shared vertical line starts at the time of the earliest created social object, records the time points of all responding events with respect to all social objects included in this thread cluster, and ends at the time of the last responding event. Fig. 6 (a) shows an example of thread cluster which includes the threads shown in Fig. 4(a).

In the responder view, we visualize the thread cluster by adding curved lines inside a crescent shape. The arc of the crescent shape represents the overall time span of all activity threads included in this cluster, and each of the curved lines represents how the particular thread spans relative to the overall time span. The vertical line is attached with horizontal arms that point to the time points when the responder responds to the corresponding threads included in this thread cluster. The y position of each arm is determined by height of the vertical line of the corresponding thread, showing its duration. Fig. 6(b) shows an example of thread cluster which includes the threads shown in Fig. 4(b).

To detect events, we cluster activity threads by using mean shift [10], a nonparametric analysis technique that adaptively generates clusters that are always centered at the positions with highest densities in the data space. We select thread features for clustering by considering the threads' closeness on the primary timeline and their semantic similarities in content (e.g., the tweets' topics).

### 4.4 System Interface and Interactions

We implement Episogram as a Web application. The system interface (as shown in Fig. 1) consists of four components, including (a) a toolbar, (b) the main display, (c) a legend, and (d) an actor list, with the following interaction support:

**Data selection.** Users can select different datasets via a dropdown menu and select one or more actors to be visualized from the actor list.

**Switching of the views.** From the toolbar, users can select different views (initiator vs. responder) to visualize the selected actors.

**Thread Aggregation.** When the data are densely distributed over time, users can aggregate the threads by automatic event detection, by selection, or by the categorical attributes associated with the corresponding social objects.

**Focusing.** Users can zoom into a particular time period by selecting a range on the time axis shown at the top of the main display, or they can select a thread to be the focus by clicking on the thread. The focused thread will be highlighted with others shown in grey.

## 5 Case Study

In this section, we illustrate how our design can be used to explore and identify patterns from social interaction data. We use two datasets that capture social interactions in different contexts: The first dataset consists of Twitter users' interactions around political debates through posting tweets and retweets, and the second dataset consists of academic publications in Physics journals which captures scholars' interactions in terms of publishing and citing papers.

### 5.1 Detecting Anomalous Behaviors in Twitter

The Twitter dataset [11] was collected during the U.S presidential election debates held in October 2012. For demonstration purposes, we selected a set of most active users who posted or retweeted the most in the data.

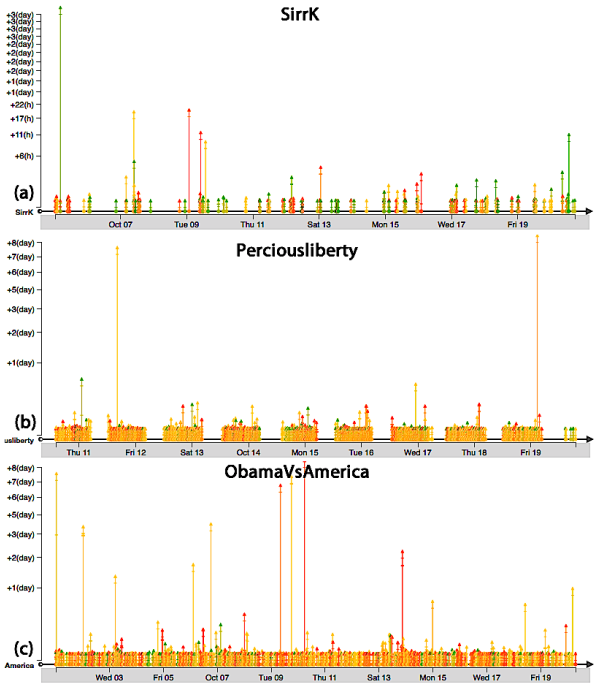


Fig. 7. Summarizing and comparing twitter users' posting behaviors in Episogram based on initiator view. (a) a typical posting behavior. (b) a periodical posting behavior. (c) a continuous posting behavior.

Fig. 7 shows the initiator view of three selected users with different posting behaviors. Most users in the dataset exhibit scattered events similar to the user "SirrK" (Fig. 7(a)) whose tweets were posted at different times across the data period and some of the tweets received more retweets than others. In this figure, the activity threads are colored based on the sentiments of the corresponding tweets (red: negative; yellow: neutral; green: positive). The users "Perciousliberty" and "ObamaVSAmerica" exhibit different patterns from those of typical users (Fig. 7(b)(c)). In particular, "Perciousliberty" (Fig. 7(b)) posted large amount of tweets regularly during a particular time period of each day. "ObamaVSAmerica" (Fig. 7(c)) posted enormous tweets incessantly throughout the entire data period. Negative sentiments are pervasive in these tweets, which can be observed from the activity threads colored in red. We have found most of these tweets express sentiments against Obama administration by reading the content of the tweets posted by the two users. Interestingly, such strong and persistent "attacks" in Twitter communication can be easily identified by visualizing the temporal patterns of posting events.

Fig. 8 shows the responder view of two users. The primary timeline records the time when the selected user retweeted other user's tweet, and the activity threads show how early or late the selected user's retweeting time was compared to other retweeting users' with respect to the same tweets. The user "CWade91" (Fig. 8(b)) tended to retweet others' tweets immediately after the tweets were

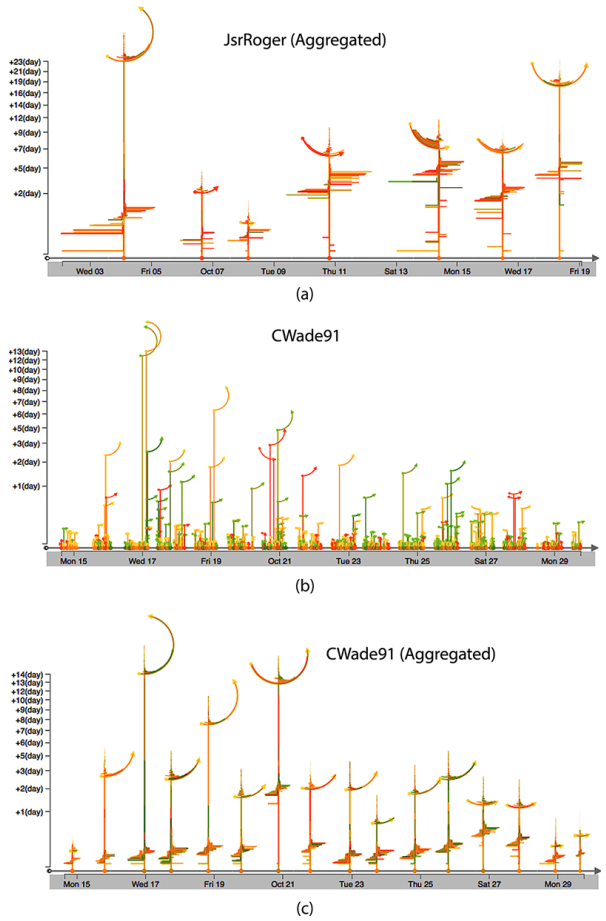


Fig. 8. Summarizing and comparing twitter users' retweeting behavors in Episogram based on responder view. (a) a typical retweeting behavior. (b) a "monitoring" behavior. (c) the aggregation view of (b).

posted (the vertical lines of these activity threads intersect mostly with the beginning of the crescent shapes). This feature can be identified more clearly by using thread aggregation function (Fig. 8(c)) which displays clusters of threads when the retweeting events occur closely in time. This early "retweeting tendency" suggests that the user is an active information spreader in Twitter. In comparison, the user "JsrRoger" (Fig. 8(a)) exhibits a more typical responding pattern – his/her retweeting to a tweet of interest may be earlier or later than other users to the same tweet.

### 5.2 Visualizing Researchers' Career Path

The publication dataset is a complete collection of papers published by Physical Review, as well as citations among them. It covers papers published in different journals such as *Physical Review (PR)*, *Physical Review Letters (PRL)*, *Reviews of Modern Physics (RMP)*, as well as *Physical Review (PR) A, B, C, D, E*, each focusing on a specific direction in physics. We selected a set of scientists as exemplary cases, who are mostly Nobel Laureates or major prize/medal awardees. All their papers, references, and citations are included for demonstration purposes.

When visualizing publication data in Episogram, the initiator view illustrates a researcher's productivity over time as well as his/her research impact generated by these publications. Each thread centered around a paper published by the researcher, indicating how the paper was cited by others over time. The responder view, on the other hand, visualizes the way in which the papers by this researcher cited existing studies. Each thread shows in an aggregated fashion, representing how a paper by the researcher cited other existing papers. Each of the cited papers is represented as an arc in the aggregated thread. In both views, the threads are colored by the journals in which the threads' corresponding papers were published. Using this encoding scheme, we demonstrate the Episogram's power of interpreting a researcher's career path.

We take professor H. Eugene Stanley as an exemplary case for our study. He is an American physicist who has made many seminal contributions to several topics of statistical physics, and was awarded the Boltzmann Medal for his contributions to phase transitions.

The first glance of Prof. Stanley's career (Fig. 9(a)) makes two impressions: (1) it is immediately clear that Prof. Stanley has been highly productive throughout his career, represented by high intensity of vertical bars over time; (2) his publications, as well as citations to these publications, are characterized by a mix of different colors: Blue corresponds to papers in premier physics journals which cover all areas of physics (Dark blue: *PR*, light blue: *PRL*), while green and red correspond to journals specializing in a particular domain of physics (*PRB* (green) covering condensed matter physics, *PRE* (pink) for statistical physics and interdisciplinary physics, and *PRA* (red) for atomic, molecular, and optical physics). Hence the mix of blue with other colors indicates publications in both premier journals that are of interest to different domains of physics as well as papers specializing in a particular field. We also observe a general shift in color from green to red/pink over time, documenting changes in research topics along his career.

More precisely, at the beginning of Prof. Stanley's career, he published most of his papers in *PRL*, a high impact premier journal which covers all topics of physics. The primary color (green) of the citations to these papers, indicating their fundamental impact to condensed matter physics. From 1971 to 1976, Prof. Stanley was extremely productive, and the high intensity of green bars during the period indicates extensive publications by him on condensed matter physics (Fig. 9(b)). The height of these bars indicates the high impact of these papers. The dense horizontal green bars in each thread, signaling his papers made significant advances within the research field. The next two decades following this significant burst of publications mark a gradual shift in his research focus. With colors shifting from green to red (Fig. 9(c)), Episogram demonstrates an increasing focus on atomic and molecular physics as well as statistical physics in his research agenda. In this period, his publications represent a great mix of papers in light blue together with green and red. Such mix indicates his research covers both papers in *PRL* that are general to all areas of physics and require more rapid dissemination and more detailed papers that impact a specific domain.

Episogram also reflects historical changes in scientific publications. From 1990 to 1993 there was a gradual split of *PRA* into two journals, *PRA* and *PRE*, with *PRE* focusing on Statistical Physics, Plasmas, Fluids, and Related Inter-

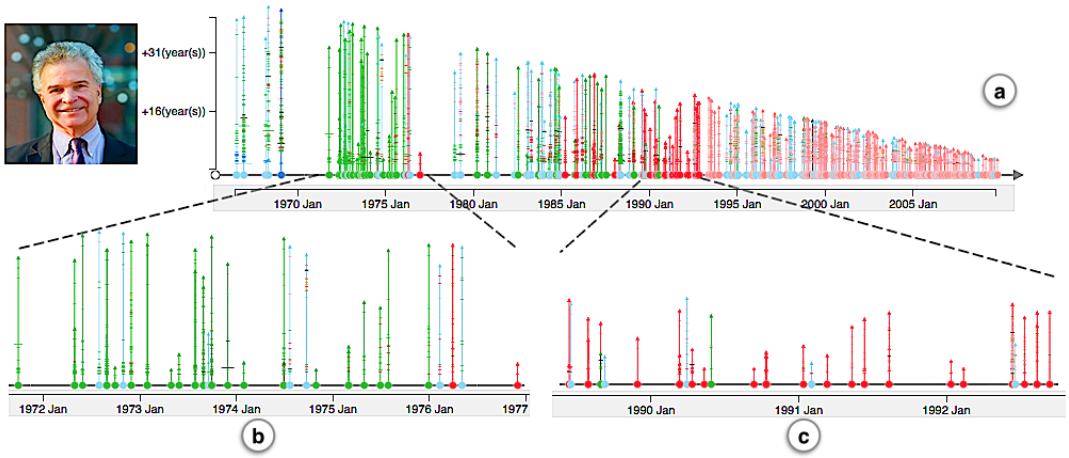


Fig. 9. Summarizing a researcher's career path in the initiator view. (a) The overview of the publication records of a processor. (b) The time period in which the professor was very productive in condensed matter physics. (c) The time period in which the professor focused on atomic, molecular, and optical physics.

disciplinary Topics. Clearly, Prof. Stanley's research is very related to the focus of *PRE*, and we observe an interesting change of colors from red to pink following this split of journals. In addition, we also observe a general decrease in the height of vertical bars. Hence more recent papers have less time to accumulate their citations.

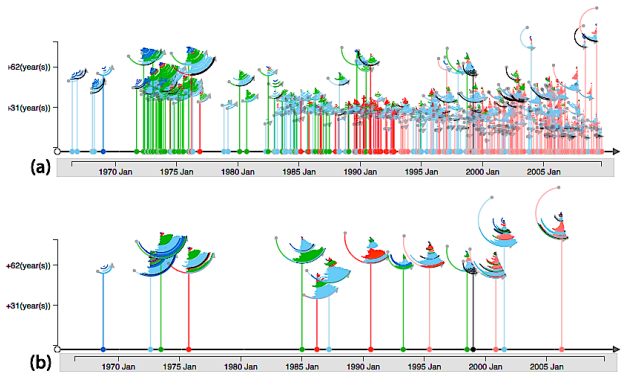


Fig. 10. Visualization of a researcher's publications in responder view: (a) threads of individual publications, and (b) thread aggregation.

The aggregated responder view of the same data (Fig. 10(a)) provides us another perspective of Prof.Stanley's career based on the way he references other papers. At the early stage of his career, he mostly cited the latest papers in his publications, demonstrating him as an early adopter of new ideas, partially explaining the observed impact of his work. At a later stage, especially after 1995, he cited a higher fraction of older papers in his publications. This pattern is potentially due to a combination of two factors, including the temporal cutoff of our dataset in 2009 and his increasing focus on well-known or longstanding problems in his research. The aggregated threads computed by mean-shift bring more visual clarity for the observed patterns (Fig. 10(b)).

## 6 User Study and Discussion

We conduct a controlled within subject study to compare Episogram with the traditional timeline view, Gantt Chart, based on a set of pattern exploration tasks[2]. The study results suggested the design effectively conveyed both detailed and overall pictures of different actors social interactions, through assisting users to identify data elements in the elementary tasks, to identify interaction patterns in the synoptic tasks, and to characterize actor interaction tendency in the higher-level synoptic tasks.

Particularly, our design complements the existing network representations by offering the capacity of summarizing interaction history that facilitates an understanding of how individual actors act and react as part of the larger network. When compared with traditional timeline views, Episogram has many key features. The timeline view resembles typical compound-event based timeline design such as Gantt chart in which the primary timeline and the activity threads share a common time axis. However, in this layout, the activity threads and primary timeline may be displayed far apart when data increase, making identifying and comparing patterns difficult. On the other hand, Episogram directly connects the primary timeline with activity threads, providing clearly the context of an interaction event and its subsequent events. Each thread is displayed with length encoding the relative duration of the thread – despite its limitation of conveying exact temporal information, the design decision was made for easily comparing the temporal relationship of the interaction event initiated or responded by the actors of interest.

In addition to the controlled user study, we also interviewed two expert users from different but related disciplines. The first expert is a Ph.D. candidate in Applied Mathematics and Computer Science from a European university with expertise in social networks and human mobility. The second expert is a postdoctoral fellow in Physics from the United States with expertise in network science. Both experts have published extensively on social network analysis, and they are rather familiar with the publication datasets used in our study. Both experts were very much impressed by the rich information offered by Episogram

2. Referring to supplemental materials for details about the studies and interviews: http://nancao.org/pubs/cao_cga_episogram_si.pdf.

as well as the design itself. The first expert particularly appreciated that Episogram sophisticatedly translates the citation statistics into visual patterns: "First time you get to look at these patterns!" The second expert highlighted the utility of our tool by comparing our tool with the simple or aggregated charts provided in citation search engine such as Google Scholar[3]. She pointed out that one novel aspect of our tool is to allow viewing how a scholar was cited by others in the absolute and relative temporal dimensions, and thus "[we can] have all scientists' productivity at a glance". Both experts agreed independently on the most useful and interesting feature offered by our tool is the aggregation function, e.g., papers or citations can be aggregated by similarity and still differentiated by their published journals and believe "it is a useful approach for reducing the clutter."

From the above studies and interviews, we also note that our design has some limitations mentioned by our users and experts: (1) Lack of network overview: the egocentric design does not allow for viewing all interactions between any two actors in a social network. We believe this limitation can be addressed by integrating our current design with a typical node-link network representation. (2) Overplotting: the rich patterns provided in the activity threads can be overwhelming if the selected actor was very active or productive. In real-world dataset, the chance of seeing the cluttered activities for an actor is rare due to the well-known power-law phenomena [12]. On the other hand, when users are interested in visualizing actors with many activities, there are several ways to effectively reduce the visual clutters: (a) users can *select* activities by categorical attributes; (b) users can *zoom in* to a particular time period; (c) users can *aggregate* activities using the aggregation function. We believe these additional tools help balance the richness and clarity in our original visual design.

## 7 Conclusion

In this paper, we presented Episogram, an interactive visualization for exploring and summarizing social interaction data. Our proposed visualization was designed based on an anatomy of social interactions in which the actors and social objects involved in the social interactions can be formally represented as a time-varying tripartite network. Particularly, a social interaction process is visualized through displaying and aggregating such tripartite networks along multiple temporal dimensions, from different actors' egocentric perspectives. This design aims to assist in a variety of user tasks ranging from elementary tasks to higher-level pattern discovery. It allows users to generate multiple views for different actors' social interaction history and to compare multiple actors in an integrated display. Our evaluation, including case studies and controlled user study, demonstrates the usefulness of Episogram.

Our future work includes two directions: (1) conduct user studies to evaluate the scalability of our visual designs; (2) develop a visual analysis systems for detecting, analyzing and visualizing different user behaviors via Episogram and other types of visualizations such as node-link graphs. We will also apply this system to analyze other datasets such as email archives.

3. http://scholar.google.com/

## Acknowledgments

This work is partially sponsored by the U.S. Defense Advanced Research Projects Agency (DARPA) under the Social Media in Strategic Communication (SMISC) program (Agreement Number: W911NF-12-C-0028). DW is supported by Air Force Office of Scientific Research under agreement number FA9550-15-1-0162. The authors wish to thank P. Deville and R. Sinatra for providing us with author disambiguated datasets as well as many useful discussions along the way.

## References

[1] P. Dourish and V. Bellotti, "Awareness and coordination in shared workspaces," in *Proceedings of the 1992 ACM conference on Computer-supported cooperative work*. ACM, 1992, p. 107114. [Online]. Available: http://dl.acm.org/citation.cfm?id=143468

[2] S. Miksch and H. Schumann, *Visualization of time-oriented data, Chapter 3*. Springerverlag London Limited, 2011.

[3] M. Ogawa, K.-L. Ma, C. Bird, P. Devanbu, and A. Gourley, "Visualizing social interaction in open source software projects," in *Visualization, 2007. APVIS'07. 2007 6th International Asia-Pacific Symposium on*. IEEE, 2007, p. 2532. [Online]. Available: http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=4126214

[4] P. Riehmann, M. Hanfler, and B. Froehlich, "Interactive sankey diagrams," in *IEEE Symposium on Information Visualization*. IEEE, 2005, pp. 233–240.

[5] R. F. Erbacher, K. L. Walker, and D. A. Frincke, "Intrusion and misuse detection in large-scale systems," *IEEE Computer Graphics and Applications*, vol. 22, no. 1, pp. 38–47, 2002.

[6] B. J. Fry, "Organic information design," Ph.D. dissertation, Massachusetts Institute of Technology, 2000.

[7] R. Xiong and J. Donath, "Peoplegarden: creating data portraits for users," in *Proceedings of the ACM symposium on User interface software and technology*. ACM, 1999, pp. 37–44.

[8] F. B. Viégas, M. Wattenberg, and K. Dave, "Studying cooperation and conflict between authors with history flow visualizations," in *Proceedings of the SIGCHI conference on Human factors in computing systems*. ACM, 2004, pp. 575–582.

[9] N. Simon, *The Participatory Museum*. Museum 2.0, 2010.

[10] Y. Cheng, "Mean shift, mode seeking, and clustering," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 17, no. 8, pp. 790–799, 1995.

[11] Y.-R. Lin, B. Keegan, D. Margolin, and D. Lazer, "Rising tides or rising stars?: Dynamics of shared attention on twitter during media events," *PloS one*, vol. 9, no. 5, p. e94093, 2014.

[12] M. E. Newman, "The structure and function of complex networks," *SIAM review*, vol. 45, no. 2, p. 167256, 2003. [Online]. Available: http://epubs.siam.org/doi/abs/10.1137/S003614450342480

**Nan Cao** is a Research Staff Member at IBM T. J. Watson Research Center. His research interests include data visualization, visual analysis, and data science. He creates novel visualizations for representing complex (i.e., big, dynamic, multivariate, heterogeneous, and multirelational) graph data in the domains of social science and medical informatics.

**Yu-Ru Lin** is an assistant professor at the School of Information Sciences, University of Pittsburgh. Her research interests include human mobility, social and political network dynamics, and computational social science. She has developed computational approaches for mining and visualizing large-scale, time-varying, heterogeneous, multi-relational, and semi-structured data.

**Fan Du** is a computer science Ph.D. student at the University of Maryland, College Park. His research focuses on data visualization and human-computer interaction, especially on analyzing healthcare data and user activity logs. He received his bachelor's degree from Zhejiang University with honors.

**Dashun Wang** is Assistant Professor of Information Sciences and Technology at the Pennsylvania State University. Through the lens of large-scale datasets, his work focuses on using and developing tools of network science to help improve the way in which we understand complexity and discover the underlying principles governing self-organized systems. His work has been applied to understand and predict social interactions, human mobility, knowledge production and scientific impact, and has been featured in Nature, Science, MIT Technology Review, The Economist, The Boston Globe, Physics World, among other outlets.